A note on the Lasso in Model Selection

Yizao Wang
Ecole Polytechnique

March 29, 2007

1 Introduction

Lasso, in short for "Least Absolute Shrinkage and Selection Operator", is first proposed by Tibshirani in 1994 [3]. This is a method for estimation in linear models. The related procedures are for example LARS (least angle regression) and FSW (forward stagewise regression). In fact, in [1], Efron et al. show that LARS can be seen a generalization of FSW and Lasso. These methods are widely used in different domains.

This report recovers the important results of the technical report [2], where Leng et. al show one limit of the Lasso and the related procedures. When the Lasso is tuned with prediction accuracy, given a model with noisy components and under certain weak condition, the probability that the method does not select the correct model is strictly bigger than a positive constant, independent of the size of sample. The authors give a simple example for the 2d case, and in general, the statement holds for the model in high dimension with orthonormal design matrix. All the results are intuitively not surprising. Moreover, as proclaimed by the authors, these should not be interpreted as a criticism of the Lasso and the related methods as variable selection tools.

The proofs given in this report are different from the ones in the original technical report. Moreover, we will see that in the original report the 2d case is based on a questionable expression of the lasso solution.

2 Lasso

We consider the common Gaussian linear regression model

$$\mathbf{y} = X\beta + \epsilon\,,$$

where $\mathbf{y} = (y_1, \cdots, y_n)^\top$ are the responses, $\beta = (\beta_1, \cdots, \beta_d)^\top$ are the regression coefficients, $X = (\mathbf{x_1}, \cdots, \mathbf{x_d})$ is the covariate matrix, and $\epsilon = (\epsilon_1, \cdots, \epsilon_n) \sim N(0, \sigma^2 I_n)$ are the normal noises. Without loss of generality, throughout this article we assume that the covariates have been standardized to mean 0 and variance 1, and the response has mean 0. That is,

$$\mathbf{1}^\top \mathbf{y} = 0, \quad \mathbf{1}^\top \mathbf{x_j} = 0, \text{ and } \mathbf{x_j}^\top \mathbf{x_j} = 1 \text{ for } j = 1, \cdots, d.$$

The Lasso estimate is the solution to

$$\min_\beta (\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta), \quad \text{s.t.} \quad \sum_{j=1}^{d} |\beta_j| \le t \tag{1}$$

An alternative formulation of the Lasso is to solve the penalized likelihood problem

$$\min_{\beta} \frac{1}{n}(\mathbf{y} - X\beta)^\top(\mathbf{y} - X\beta) + \lambda \sum_{j=1}^{d} |\beta_j| \,. \tag{2}$$

(1) and (2) are equivalent in the sense that for any given $\lambda \in [0, +\infty)$, there exists a $t \geq 0$ such that the two problems have the same solution, and vice versa. We also note $\hat{\beta}^0$ as the ordinary least square (OLS) estimate, i.e., the solution to

$$\min_{\beta}(\mathbf{y} - X\beta)^\top(\mathbf{y} - X\beta) \,. \tag{3}$$

In the follows of this report we note β the true model and $\hat{\beta}$ the one found by Lasso (the solution of (1)).

Remark 1 *We find the notions are sometimes confusing. However we keep the notions as in [3] and [2]. To summarize, β^0 is the true model, $\hat{\beta}^0$ is the OLS solution and $\hat{\beta}$ is the solution found by Lasso.*

Most of the analysis depends on the relationship between $\hat{\beta}^0$, β^0, $\hat{\beta}^0$ and γ. In fact, when the design matrix is orthonormal, i.e. $X^\top X = I_d$, the Lasso solution has the form

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \quad j = 1, \cdots, d, \tag{4}$$

where $\gamma = \lambda/2$ for λ in (2). $(x)^+ = x, x > 0; 0, x \leq 0$. The proof is in Appendix A.1. In implementations of the Lasso (and the other methods), usually the methods is tuned with prediction accuracy, which is in terms of the squared loss (SL). Given a γ, we have an estimate $\hat{\eta} = X\hat{\beta}$ and the squared loss

$$SL(\gamma) = SL(\hat{\eta}) = (\hat{\eta} - \eta)^\top(\hat{\eta} - \eta) = (\hat{\beta} - \beta)^\top X^\top X(\hat{\beta} - \beta) \,. \tag{5}$$

Remark 2 *In practice, β is always unknown, several methods, are used to for the purpose of minimizing the squared error. However, in the original technical report, all the investigation is based on (5). I find this rather strange.*

We also need in the following analysis $\hat{\delta}$ defined by $\hat{\delta} = \hat{\beta}_0 - \beta_0$. Since have

$$\begin{aligned} \hat{\beta}^0 &= (X^\top X)^{-1} X^\top \mathbf{y} \\ &= (X^\top X)^{-1} X^\top (X\beta^0 + \epsilon) \\ &= \beta^0 + \Sigma^{-1} X^\top \epsilon \,, \end{aligned}$$

then

$$\hat{\delta} \sim N(0, \sigma^2 \Sigma^{-1}) \,. \tag{6}$$

3 A simple example

In this section, we shown that given a simple example, Lasso, when tuned to minimize the squared error, misses the right model with a certain probability. The simple example is with the true coefficient vector $\beta^0 = (\beta_1^0, 0, \cdots, 0)^\top$ where $\beta_1^0 > 0$. First we see when the dimension n is 2. Then we discuss higher dimension. In the latter case we suppose the design matrix is orthonormal, i.e. $X^\top X = \Sigma = \mathbf{I_d}$. In fact, it is almost obvious to see that Lasso misses the true model with a certain probability. However, we go through the mathematical details in this section to have a precise description. In fact, the following proofs are simpler than the ones in the original article.

3.1 2D case

We restate the notation. The true coefficient vector is $\beta^0 = (\beta_1^0, 0)^\top$, $\beta_1^0 > 0$. The standardized design matrix X is

$$X^\top X = \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

with $|\rho| < 1$. The component $\mathbf{x}_2$ is the noisy component. Note the ordinary least squares solutions by $\hat{\beta}^0$. The original report [2] said that, given $\hat{\beta}^0$, the solution $\hat{\beta}$ to the Lasso problem when $d = 2$ can be easily seen as

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \tilde{\gamma})^+, \quad j = 1, 2 \tag{7}$$

However, they did not show how they had (7) and the computation in Appendix A.2 shows in (7) $\tilde{\gamma}$ depends on ρ, $\hat{\beta}$ and $\gamma = \frac{\lambda}{2}$.

Remark 3 *In fact, in both [2] and [3], the 2D case is mentioned without calculations (it is both said easy to have). In [3] the author emphasized the condition that $\hat{\beta}_1^0 > 0, \hat{\beta}_2^0$, but this is just the case in Appendix A.2. We can still do some similar analysis as in [2], but the method is too technique with little interest and can not be generalized to higher dimension. We just ignore this part. For $\rho = 0$, we see in Section 3.2 a general proof.*

3.2 Simple case in high dimension

In this section we generalize the case in high dimension. We add a new condition on the model, that is the design matrix is strictly definite positive. We prove the following lemma, which is stronger than in the original paper [2] where the design matrix is supposed to be orthonormal.

Lemma 1 *When $\beta^0 = (\beta_1^0, 0, \cdots, 0)^\top$ with $(d-1) > 0$ zero components and $X^\top X = I_d$, the Lasso tuned with prediction accuracy selects the right model if and only if $\hat{\delta} = \hat{\beta}^0 - \beta^0 \in \mathcal{R}$, where*

$$\mathcal{R} = \left\{ \delta \in \mathbf{R}^d : \delta_1 \beta_1^0 > 0, \quad |\delta_1| > \max\{|\delta_2|, \cdots, |\delta_d|\} \right\},$$

that is, the probability of the right model being selected is $1/(2d)$.

Proof: Recall the form of the Lasso solution:

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \quad j = 1, \cdots, d\,.$$

If the correct model is selected, i.e. $\hat{\beta} = \beta^0$, we need

$$|\hat{\beta}_1^0| - \max\{|\hat{\beta}_2^0|, \cdots, |\hat{\beta}_d^0|\} \geq |\beta_1^0|\,.$$

Without loss of generality, assume $\beta_1^0 > 0$ and $|\hat{\delta}_2| = \max\{|\hat{\delta}_2|, \cdots, |\hat{\delta}_d|\}$. (We have in fact $\hat{\beta}_j^0 = \hat{\delta}_j, j = 2, \cdots, d$.) The first remark is $\hat{\beta}_1^0 < 0$ implies $\hat{\beta}_1 \leq 0 < \beta_1^0$ and the Lasso never selects the true model. Then when $\hat{\beta}_1^0 \geq 0$, to have $\hat{\beta} = \beta^0$ we need

$$\hat{\beta}_1^0 - |\hat{\beta}_2^0| \geq \beta_1^0 \Leftrightarrow \hat{\delta}_1 > |\hat{\delta}_2|\,.$$

Then $\hat{\delta}_1 \geq |\hat{\delta}_2|$ is a necessary condition for the correct model being selected. It is in fact also a sufficient condition, since in this case,

$$\begin{aligned} SL(\gamma) &= (\hat{\beta}_1 - \beta_1^0)^2 + \sum_{i=2}^{d} (\hat{\beta}_2)^2 \\ &= \begin{cases} (\beta_0)^2 & \text{if } \gamma \geq |\hat{\beta}_1^0| \\ (\hat{\delta}_1 - \gamma)^2 & \text{if } |\hat{\beta}_2^0| \leq \gamma < \hat{\beta}_1^0 \\ (\hat{\delta}_1 - \gamma)^2 + \sum_{i=2}^{d} (\hat{\beta}_2)^2 & \text{if } \gamma < |\hat{\beta}_2^0| \end{cases} , \end{aligned}$$

Again we have $\text{argmin} SL(\gamma) = \hat{\delta}_1$, which gives $\hat{\beta} = \beta_0$. The correct model is selected when $\hat{\delta}$ belongs to:

$$\mathcal{R} = \left\{ \delta \in \mathbf{R}^d; \delta_1 > \max\{|\delta_2|, \cdots, |\delta_d|\} \right\}$$

We proved that the probability of selecting the correct model is $1/2d$. ∎

4 Higher dimensional problems with orthogonal designs

Theorem 1 *When the true coefficient vector is $\beta^0 = (\alpha_1, \cdots, \alpha_{d_1}, 0, \cdots, 0)^\top$ with $d_2 = d - d_1 > 0$ zero coefficients and $X^\top X = I_d$, if the Lasso is tuned according to prediction accuracy, then it selects the right model with probability 0 when $d_1 > 1$.*

Proof: Again, when $X^\top X = I_d$, the Lasso solution is as in (4). Then first we need

$$|\hat{\beta}_j^0| > |\hat{\beta}_k^0|, \text{ for } j \in \{1, \cdots, d_1\} \text{ and } k \in \{d_1 + 1, \cdots, d\} . \tag{8}$$

With (8), if Lasso is tuned at γ^*, then it is necessary to have

$$\min\{|\hat{\beta}_1^0|, \cdots, |\hat{\beta}_{d_1}^0|\} > \gamma^* > \max\{|\hat{\beta}_{d_1+1}^0|, \cdots, |\hat{\beta}_d^0|\} \tag{9}$$

and

$$\alpha_i = \text{sign}(\hat{\beta}_i^0)(|\hat{\beta}_i^0| - \gamma^*) . \tag{10}$$

Since (10) if valid for all $i = 1, \cdots, d_1$, we need

$$\text{sign}(\alpha_i) = \text{sign}(\hat{\beta}_i^0), \ |\hat{\beta}_i^0 - \alpha_i| = C \geq 0, \text{ for } i = 1, \cdots, d_1 . \tag{11}$$

Obviously here $C \geq \gamma^* \geq 0$. With the same notation,

$$\hat{\delta}_i = \hat{\beta}_i^0 - \alpha_i^0 = \text{sign}(\alpha_i) \cdot C \text{ for } i = 1, \cdots, d_1 ,$$

and $\hat{\delta}_i = \hat{\beta}_i^0$ for $i > d_1$. Then, by a similar analysis as before, we see that Lasso tuned to prediction accuracy selects the right model if and only if $\hat{\delta}$ belongs to

$$\mathcal{R} = \left\{ \delta \in \mathbf{R}^d : \ \delta_i = \text{sign}(\alpha_i) \cdot C, C \geq \max\{|\delta_{d_1+1}^0|, \cdots, |\delta_d^0|\}, \text{ for } i = 1, \cdots, d_1 \right\} .$$

We finally proved that the Lasso tuned to prediction accuracy selects the right model with probability 0 when $d_1 > 1$. ∎

5 Discussion and conclusion

- In practice prediction accuracy is the golden standard and the Lasso can improve greatly over the ordinary least estimate in terms of accuracy.
- The authors intended to emphasize that 'choosing the correct model' might not be a good criteria for Lasso and the related methods since there are the cases when Lasso can never find the true coefficients. This is not saying that Lasso is not a good method. Other methods exist for consistent variable selection.
- [1] shows the relation between Lasso, Lars and FSW. Then the conclusions we have for Lasso are valuable for Lars and FSW methods, with eventually slight changement on the conditions.
- However, all the analysis is based on the prediction error formalized by (5). This is somehow questionable since in practice the λ is never selected as in the analysis above.
- The proofs in [2] are not optimized. The analysis for simple example in 2D is even questionable. The ones in this report are clearer and easier.

A The form of the Lasso solution

A.1 With orthonormal design matrix

In this section, we prove that when $X^\top X = I_d$, the form of the Lasso solution to the problem

$$\min_\beta (\mathbf{y} - X\beta)^\top (\mathbf{y} - X\beta) + \lambda \sum_{i=1}^{d} |\beta_i|$$

is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \quad j = 1, \cdots, d, \tag{12}$$

where $\hat{\beta}^0$ is the ordinary least square estimate.

Proof: First the problem is equivalent to

$$\begin{aligned} & \min_\beta -2\mathbf{y}^\top X\beta + \beta^\top \beta + \lambda \sum_{i=1}^{d} |\beta_i| \\ = & \min_\beta -2\hat{\beta}^0 \beta + \beta^\top \beta + \lambda \sum_{i=1}^{d} |\beta_i| \\ = & \min_\beta \sum_{i=1}^{d} \left(-2\hat{\beta}_i^0 \beta_i + \beta_i^2 + \lambda |\beta_i| \right) . \end{aligned}$$

Then, we need only to deal with each index i separately. To simplify we erase the index in the following. Then we are going to solve

$$\min \left\{ \min_{\beta \geq 0} (-2\hat{\beta}^0 \beta + \beta^2 - \lambda\beta), \min_{\beta \leq 0} (-2\hat{\beta}^0 \beta + \beta^2 - \lambda\beta) \right\} .$$

Respectively we have:

$$\min_{\beta\geq 0}(-2\hat{\beta}^0\beta+\beta^2-\lambda\beta)=\begin{cases}-(\hat{\beta}^0-\frac{\lambda}{2})^2 & \text{with } \beta^*=\hat{\beta}^0-\frac{\lambda}{2} \quad \text{if } \hat{\beta}-\frac{\lambda}{2}>0\\ 0 & \text{with } \beta^*=0 \quad \text{if } \hat{\beta}-\frac{\lambda}{2}\leq 0\end{cases}$$

$$\min_{\beta\geq 0}(-2\hat{\beta}^0\beta+\beta^2+\lambda\beta)=\begin{cases}-(\hat{\beta}^0+\frac{\lambda}{2})^2 & \text{with } \beta^*=\hat{\beta}^0+\frac{\lambda}{2} \quad \text{if } \hat{\beta}+\frac{\lambda}{2}>0\\ 0 & \text{with } \beta^*=0 \quad \text{if } \hat{\beta}+\frac{\lambda}{2}\leq 0\end{cases}$$

To summarize, replacing $\frac{\lambda}{2}$ by γ, we recover the expression of the minimizer $\hat{\beta}^*$ as in (12). ∎

A.2 General 2D case

Suppose the design matrix X is

$$X^\top X=\Sigma=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix},$$

with $|\rho|<1$. Denote $\gamma=\frac{\lambda}{2}$. Remember $\hat{\beta}^0=(X^\top X)^{-1}X^\top\mathbf{y}=\Sigma^{-1}X^\top\mathbf{y}$. Then $\mathbf{y}^\top X\beta=(\hat{\beta}^0)^\top\Sigma\beta$ and the Lasso problem becomes:

$$\begin{aligned}\min_\beta P(\beta) &= \min_\beta\Big\{(\beta-2\hat{\beta}^0)^\top\Sigma\beta+2\gamma(|\beta_1|+|\beta_2|)\Big\}\\ &= \min_\beta\Big\{\beta_1^2+2\rho\beta_1\beta_2+\beta_2^2-2\beta_1(\hat{\beta}_1^0+\rho\hat{\beta}_2^0-\text{sign}(\beta_1)\cdot\gamma)\\ &\qquad -2\beta_2(\rho\hat{\beta}_1^0+\hat{\beta}_2^0-\text{sign}(\beta_2)\cdot\gamma)\Big\}\end{aligned}$$

Discuss respectively according to the signs of β_1 and β_2. To simplify we suppose that $\hat{\beta}_1^0>0,\hat{\beta}_2^0>0$.

1) $\beta_1>0,\beta_2>0$. Then

$$\begin{aligned}\frac{\partial P(\beta)}{\partial\beta_1} &= 2\beta_1+2\rho\beta_2-2(\hat{\beta}_1^0+\rho\hat{\beta}_2^0-\gamma)=0\,,\\ \frac{\partial P(\beta)}{\partial\beta_2} &= 2\beta_2+2\rho\beta_1-2(\hat{\beta}_2^0+\rho\hat{\beta}_1^0-\gamma)=0\,.\end{aligned}$$

$$\Leftrightarrow\quad\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}=\begin{pmatrix}\hat{\beta}_1^0+\rho\hat{\beta}_2^0-\gamma\\ \hat{\beta}_2^0+\rho\hat{\beta}_1^0-\gamma\end{pmatrix}$$

$$\Leftrightarrow\quad\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}=\frac{1}{1-\rho^2}\begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix}\begin{pmatrix}\hat{\beta}_1^0+\rho\hat{\beta}_2^0-\gamma\\ \hat{\beta}_2^0+\rho\hat{\beta}_1^0-\gamma\end{pmatrix}=\begin{pmatrix}\hat{\beta}_1^0-\frac{1}{1+\rho}\gamma\\ \hat{\beta}_2^0-\frac{1}{1+\rho}\gamma\end{pmatrix}$$

We note the solution for $\beta_1>0,\beta_2>0$ by

$$\begin{cases}\beta_1^* &= (\hat{\beta}_1^0-\frac{1}{1+\rho}\gamma)^+\\ \beta_2^* &= (\hat{\beta}_2^0-\frac{1}{1+\rho}\gamma)^+\end{cases}.$$

Remark 4 *However, when $\beta_1^*\beta_2^* = 0$, β^* is not necessarily the minimum for $\beta_1 \geq 0, \beta_2 \geq 0$. However, we use the notation above to simplify. In fact, we need to calculate the minimum on the whole x and y axis when the minimum is not achieved beyond the axis. The minimum on x-axis, when $\beta_2 = 0$, is given at $(\hat{\beta}_1^0 + \rho\hat{\beta}_2^0 - \gamma)^+$. The minimum on y-axis is given at $(\hat{\beta}_2^0 + \rho\hat{\beta}_1^0 - \gamma)^+$. Moreover, when $\hat{\beta}_1^0 > \hat{\beta}_2^0$, the minimum on x-axis is the global minimum, otherwise the one on y-axis is.*

2) $\beta_1 < 0, \beta_2 < 0$. Then similarly

$$\left\{ \begin{array}{rcl} \beta_1^* & = & -(\hat{\beta}_1^0 + \frac{1}{1+\rho}\gamma)^- \\ \beta_2^* & = & -(\hat{\beta}_2^0 + \frac{1}{1+\rho}\gamma)^- \end{array} \right. ,$$

where $(x)^- = -x, x < 0; \quad 0,$ otherwise .

3) $\beta_1 > 0, \beta_2 < 0$. Then

$$\begin{array}{rcl} \frac{\partial P(\beta)}{\partial \beta_1} & = & 2\beta_1 + 2\rho\beta_2 - 2(\hat{\beta}_1^0 + \rho\hat{\beta}_2^0 - \gamma) = 0\,, \\ \frac{\partial P(\beta)}{\partial \beta_2} & = & 2\beta_2 + 2\rho\beta_1 - 2(\hat{\beta}_2^0 + \rho\hat{\beta}_1^0 + \gamma) = 0\,. \end{array}$$

$$\Rightarrow \left\{ \begin{array}{rcl} \beta_1^* & = & (\hat{\beta}_1^0 - \frac{1}{1-\rho}\gamma)^+ \\ \beta_2^* & = & -(\hat{\beta}_2^0 + \frac{1}{1-\rho}\gamma)^- \end{array} \right. .$$

4) $\beta_1 < 0, \beta_2 > 0$. Then similarly

$$\left\{ \begin{array}{rcl} \beta_1^* & = & -(\hat{\beta}_1^0 + \frac{1}{1-\rho}\gamma)^- \\ \beta_2^* & = & (\hat{\beta}_2^0 - \frac{1}{1-\rho}\gamma)^+ \end{array} \right. .$$

To summarize, we see that when $\hat{\beta}_1^0 > 0, \hat{\beta}_2^0 > 0$, the minimum can only be achieved in the first quadrant if not on the axis. It is the case when $\min\{\hat{\beta}_1^0, \hat{\beta}_2^0\} > \frac{1}{1+\rho}\gamma$, and we have

$$\hat{\beta}_i = \hat{\beta}_i^0 - \frac{1}{1+\rho}\gamma\,.$$

However, considering the case when the minimum is on the axis as in Remark 4, we can not unify the solution as

$$\hat{\beta}_i = (\hat{\beta}_i^0 - \frac{1}{1+\rho}\gamma)^+\,. \tag{13}$$

To see this, take for example $\hat{\beta}_1^0 = \frac{1+2\epsilon}{1+\rho}\gamma, \hat{\beta}_2^0 = \frac{1-\epsilon}{1+\rho}\gamma$. Then the minimum is on the axis at $(\frac{(2-\rho)\epsilon}{1+\rho}\gamma, 0)$, but (13) yields $(\frac{2\epsilon}{1+\rho}\gamma, 0)$. To summarize (this is the case when $\hat{\beta}^0 > 0$), we can rewrite the solution form as

$$\hat{\beta}_j = \mathrm{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \tilde{\gamma})^+, \quad j = 1, 2\,,$$

$$\tilde{\gamma} = \left\{ \begin{array}{ll} \frac{1}{1+\rho} & \text{if } \min\{\hat{\beta}_1^0, \hat{\beta}_2^0\} > \frac{1}{1+\rho}\gamma \\ \gamma - \rho\hat{\beta}_2^0 & \text{if } \hat{\beta}_1^0 > \frac{1}{1+\rho}\gamma > \hat{\beta}_2^0 > 0 \\ \gamma - \rho\hat{\beta}_1^0 & \text{if } \hat{\beta}_2^0 > \frac{1}{1+\rho}\gamma > \hat{\beta}_1^0 > 0 \\ 0 & \text{if } 0 < \max\{\hat{\beta}_1^0, \hat{\beta}_2^0\} < \frac{1}{1+\rho}\gamma \end{array} \right. ,$$

where $\gamma = \frac{\lambda}{2}$.

References

[1] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression, 2002.

[2] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. 2004.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso, 1994.